AP Statistics

—— 陈依璟

Unit 1: Exploring One-Variable Data	E	**15–23%**
Unit 2: Exploring Two-Variable Data		**5–7%** X
Unit 3: Collecting Data	D	**12–15%**
Unit 4: Probability, Random Variables, and Probability Distributions	C	**10–20%**
Unit 5: Sampling Distributions		**7–12%** X
Unit 6: Inference for Categorical Data: Proportions	B	**12–15%**
Unit 7: Inference for Quantitative Data: Means	A	**10–18%**
Unit 8: Inference for Categorical Data: Chi-Square		**2–5%** X
Unit 9: Inference for Quantitative Data: Slopes		**2–5%** X

intersection 交集　　union 并集

$$P(A|B) = \frac{P(A\cap B)}{P(B)} \qquad P(A\cap B) = P(A)\cdot P(B|A)$$

conditional probability ('given that'): $\frac{\text{两条件交集}}{\text{given that 后条件}}$

Conditional Probability Formula

$$\underbrace{P(A|B)}_{\text{Probability of } A \text{ given } B} = \frac{\overbrace{P(A\cap B)}^{\text{Probability of } A \text{ and } B}}{\underbrace{P(B)}_{\text{Probability of } B}}$$

independent

$$P(A\cap B) = P(A)\cdot P(B)$$

$$P(A|B) = P(A|\bar{B})$$

加减不影响标准差

$$Y = a + bX$$

$$\mu_Y = a + b\mu_X$$

$$\sigma_Y = |b|\sigma_X$$

两边同时平方变成方差　$\sigma_Y^2 = |b|^2\sigma_X^2$

$$\Downarrow$$

$$Var(Y) = |b|^2 Var(X)$$

X and Y are independent

if $D = X - Y$

$$E(D) = \mu_D = \mu_X - \mu_Y$$

$$\sigma_D = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

$$T = X + Y$$

$$\mu_T = \mu_X + \mu_Y$$

$$\sigma_T^2 = \sigma_X^2 + \sigma_Y^2 \Rightarrow \sigma_T = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

$$\sqrt{n\sigma_X^2 + n\sigma_Y^2}$$

$$\sqrt{n_1p_1q_1 + n_2p_2q_2}$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

e.g. $P(A): 0.8$　　$P(B): 0.7$

$$0.8 + 0.7 - 0.8 \times 0.7 = 0.94$$

线性关系变量求均值

$y = 5x + 2$

$\downarrow$

$y_1 = 5x_1 + 2$

$y_2 = 5x_2 + 2$

$y_3 = 5x_3 + 2$

$$\frac{y_1 + y_2 + y_3}{3} = \frac{5x_1 + 2 + 5x_2 + 2 + 5x_3 + 2}{3}$$

$$\bar{y} = 5\bar{x} + 2$$

$$\bar{y} = k\bar{x} + C$$

已知 σx^2

$$\sigma y^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2}{n}$$

$$= \frac{[5x_1 + 2 - (5\bar{x} + 2)]^2 + [5x_2 + 2 - (5\bar{x} + 2)]^2 + [5x_3 + 2 - (5\bar{x} + 2)]^2}{n}$$

$$= \frac{5^2 \Sigma (x_i - \bar{x})^2}{n}$$

$$= 5^2 \sigma x^2$$

$$\sigma y = 5 \sigma x$$

$$\sigma y^2 = k^2 \sigma x^2 \rightarrow \sigma y = k \sigma x$$

An analogy:

Probability: Starting with an animal, and figuring out what footprints it will make.

Statistics: seeing a footprint, and guessing the animal.

Normal distribution as approximation to binominal 正态分布近似二项分布

$\mu_x = np \qquad \sigma_x = \sqrt{np(1-p)}$

条件: $np \geq 10$, $nq \geq 10$

① 求出参数 (μ, σ)

② Norm. Cdf (连续性较正, $n \pm 0.5$)

大数定理

重复实验次数越多，事件发生的概率就越接近期望值.

二项分布

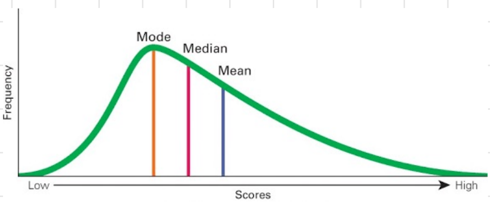
$\mu = np \qquad \sigma = \sqrt{np(1-p)}$ （当 $n = 1$ 时是伯努力分布）

几何分布

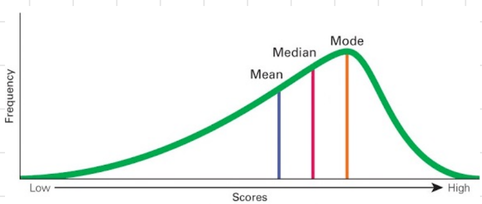
$\mu = \frac{1}{p} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$

方差: 数据的波动大小.

$$s^2 = \frac{1}{n}\left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right]$$



(a) Right-skewed distribution



(b) Left-skewed distribution

中心极限定理

$$
\begin{aligned}
E(\bar{X}) &= E\left(\frac{X_1+X_2+\cdots X_n}{n}\right) \\
&= E\left(\frac{X_1}{n}\right) + E\left(\frac{X_2}{n}\right) + \cdots + E\left(\frac{X_n}{n}\right) \\
&= \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \cdots + \frac{1}{n}E(X_n) \\
&= \frac{1}{n}\cdot\mu\cdot n \\
&= \mu
\end{aligned}
$$

$$
\begin{aligned}
Var(\bar{X}) &= Var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) \\
&= Var\left(\frac{X_1}{n}\right) + Var\left(\frac{X_2}{n}\right) + \cdots + Var\left(\frac{X_n}{n}\right) \\
&= \frac{1}{n^2}Var(X_1) + \frac{1}{n^2}Var(X_2) + \cdots + \frac{1}{n^2}Var(X_n) \\
&= \frac{1}{n^2}\cdot\sigma^2 + \frac{1}{n^2}\cdot\sigma^2 + \cdots + \frac{1}{n^2}\cdot\sigma^2 \\
&= \frac{1}{n^2}\cdot\sigma^2\cdot n \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

- 样本均值的方差是总体方差的 $\frac{1}{n}$
- 样本均值的标准差是总体标准差的 $\frac{1}{\sqrt{n}}$
- 样本均值的均值是总体的均值

如果本就服从正态分布，样本数量不用很大

$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ [比例 $\sigma_{\bar{x}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$]

统计学第一定律

Central Limit Theorem (CLT)

If $X \sim N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

If $X \sim$ any distribution with a mean μ, and variance σ^2,

then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ given that n is large.

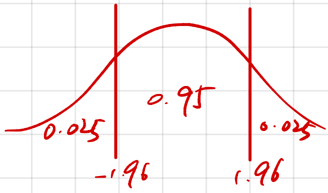
$$\boldsymbol{\sigma}_{\bar{x}} = \sqrt{\frac{\boldsymbol{\sigma}^2}{\boldsymbol{n}}} = \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$$

样本均值分布（已知总体求均值分布）

$$P(\underset{z}{-1.96} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < 1.96) = 0.95$$

$$P(-1.96\frac{\sigma}{\sqrt{n}} < \bar{x}-\mu < 1.96\frac{\sigma}{\sqrt{n}})$$

$$P(\mu-1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu+1.96\frac{\sigma}{\sqrt{n}})$$



正态分布估计标准差

$$\frac{max - min}{6} \approx \sigma$$

max: 190 min: 142

$$165 - 1.96 \times 8 < \bar{x} < 165 + 1.96 \times 8$$

$$149.32 < \bar{x} < 180.68$$

↓

95%

正态分布标准化

$$area \rightleftharpoons z = \frac{x-\mu}{\sigma}$$ 同源

marginal of error

Confidence interval

based on the Central Limit Theorem

$$P(\mu - 1.96\cdot\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96\cdot\frac{\sigma}{\sqrt{n}}) = 0.95$$ 已知总体均值估计样本均值

$$P(\underline{-1.96\cdot\frac{\sigma}{\sqrt{n}}} - \bar{x} < -\mu < \underline{1.96\cdot\frac{\sigma}{\sqrt{n}}} - \bar{x}) = 0.95$$

$$P(\underline{1.96\cdot\frac{\sigma}{\sqrt{n}}} + \bar{x} > \mu > \underline{-1.96\cdot\frac{\sigma}{\sqrt{n}}} + \bar{x}) = 0.95$$

↓

probability → $P(\bar{x} - 1.96\cdot\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\cdot\frac{\sigma}{\sqrt{n}}) = 0.95$ 已知样本均值估计总体均值

正态分布

propotion → $P(\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = 0.95$ Population proportion

比例

$\boxed{t\text{分布}\ \sigma\ \text{未知}}$ 总体标准差未知

$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

t分布

Sx是对总体 σ^2 的无偏估计

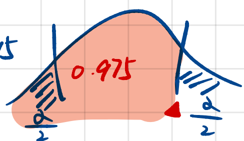
$$\parallel$$

样本30以内 $t = \frac{\bar{x}-\mu}{S_x/\sqrt{n}}$

degree freedom = sample size − 1

t分布 $p(\bar{x} - \underline{t_{\frac{\alpha}{2}, n-1}}\cdot\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2}, n-1}\cdot\frac{s}{\sqrt{n}}) = 0.95$

inverse t → critical value



standard error of the sample mean $\bar{x}$ is $\frac{s_x}{\sqrt{n}}$ (无系数)

proportion

margin of error in the confidence interval for p is $ME = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (有系数)

高成本的有把握 → n ↑, margin of error ↓, 置信区间变窄 / 低成本的无把握 n ↓

PK ①
- significance level 显著性水平 → 假设检验
- p-value 该事件和比该事件更离谱的事件概率和 → [tcdf 5, 5, 5 带入 t^*] weired

②
- $t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ 标准化 — standard error 样本均值标准差
- critical value 已标准化 [查表 inverse t]

Confidence interval

sample mean / proportion / slope ± critical value × standard error

总体均值假设检验

$n =$ 100

$\bar{x} =$ 165   s= 25 [查电子表格]

95% CI for μ:

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 165 \pm 1.984\frac{25}{\sqrt{100}} = (160.03946, 169.96054)$$ [电子表格 4, 3, 2]

$H_0 : \mu$ ≥ 170

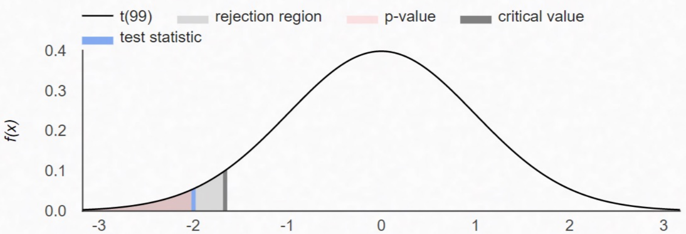
$H_a : \mu$ < 170

Significance level: $\alpha =$ 0.05 定义了小概率事件

Critical value: $-t_{\alpha,n-1} = -t_{0.05,99} = -1.660$ [inverse t]

Test statistic: $t^* = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{165-170}{25/\sqrt{100}} = -2.000$

p-value: $P(t_{(n-1)} < t^*) = P(t_{(99)} < -2.000) = 0.02412$ [电子表格 4, 4, 2]



假设检验中，Ha 决定左右尾　从单尾变双尾 p value 翻倍

双样本问题

双样本（样本均值差标准化）

$$\text{Test statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - (\Delta_0)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \Rightarrow S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$\Delta_0 \rightarrow H_0$ 的值

$S_p \downarrow$ 替代 σ $\downarrow$ σ 未知时用 S_p 代替

样本均值差的标准差（P210）

$S_p \rightarrow$ pooled 合并后的标准差

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$$

$\begin{cases} s \begin{cases} = \\ \neq \end{cases} \\ \sigma \ \text{样本方差} \end{cases}$

双样本差的 confidence level

$$\bar{x}_1 - \bar{x}_2 \pm t_{\frac{\alpha}{2},\ \underbrace{(n_1+n_2)-2}_{\text{自由度}}} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

双样本总体均值差假设检验

Statistical Inference for $\mu_1 - \mu_2$

$n_1 =$	400	$\bar{x}_1 =$	171	$\sigma_1 =$	20
$n_2 =$	360	$\bar{x}_2 =$	168	$\sigma_2 =$	16

95% CI for $\mu_1 - \mu_2$:

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 171 - 168 \pm 1.960\sqrt{\frac{20^2}{400} + \frac{16^2}{360}}$$
$$= 3.00000 \pm 2.56382$$
$$= (0.43618, 5.56382)$$

$H_0 : \mu_1 - \mu_2$ = 6

$H_a : \mu_1 - \mu_2$ ≠ 6

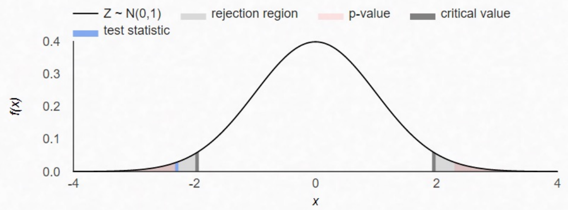
Significance level: $\alpha =$ 0.05

Critical value: $\pm z_{\alpha/2} = \pm z_{0.025} = \pm 1.960$

Test statistic:

$$z^* = \frac{(\bar{x}_1 - \bar{x}_2) - (\Delta_0)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(171-168)-(6)}{\sqrt{\frac{20^2}{400} + \frac{16^2}{360}}} = -2.293$$

p-value: $2P(Z > |z^*|) = 2P(Z > 2.293) = 0.02182$



单比例假设检验（单变量）

$$P\left(\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 0.95$$

$$t^*: z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

— H_0

Statistical Inference for p

$n =$ 100 　 $\hat{p}=$ 0.5

Inference method: Wald

95% CI for p:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.50000 \pm 1.960\sqrt{\frac{0.50000(1-0.50000)}{100}}$$
$$= 0.50000 \pm 0.09800$$
$$= (0.40200, 0.59800)$$

standard error
标准误

$H_0 : p$ = 0.6

$H_a : p$ ≠ 0.6

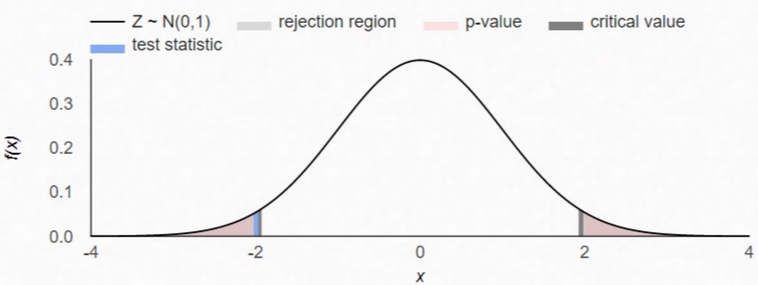
Significance level: $\alpha =$ 0.05

Critical value: $\pm z_{\alpha/2} = \pm z_{0.025} = \pm 1.960$

Test statistic:

$$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \frac{0.50000 - (0.6)}{\sqrt{\frac{0.50000(1-0.50000)}{100}}} = -2.000$$

p-value: $2P(Z > |z^*|) = 2P(Z > 2.000) = 0.04550$



双样本比例差假设检验（双变量）

Statistical Inference for $p_1 - p_2$

$n_1 =$ 400 $\quad \hat{p}_1 =$ 0.7

$n_2 =$ 600 $\quad \hat{p}_2 =$ 0.65

95% CI for $p_1 - p_2$:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= 0.7 - 0.65 \pm 1.960\sqrt{\frac{0.7(1-0.7)}{400} + \frac{0.65(1-0.65)}{600}}$$

$$= 0.05000 \pm 0.05893$$

$$= (-0.00893, 0.10893)$$

$H_0 : p_1 - p_2$ = 0.12

$H_a : p_1 - p_2$ ≠ 0.12

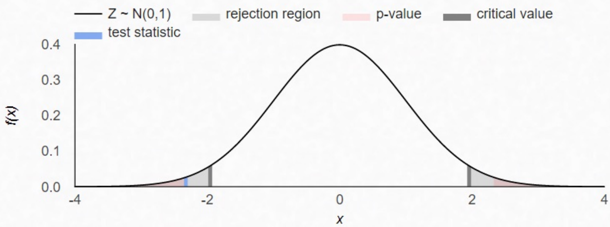
Significance level: $\alpha =$ 0.05

Critical value: $\pm z_{\alpha/2} = \pm z_{0.025} = \pm 1.960$

Test statistic:

$$z^* = \frac{(\hat{p}_1 - \hat{p}_2) - (\Delta_0)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{(0.7 - 0.65) - (0.12)}{\sqrt{\frac{0.7(1-0.7)}{400} + \frac{0.65(1-0.65)}{600}}} = -2.328$$

p-value: $2P(Z > |z^*|) = 2P(Z > 2.328) = 0.01991$



Paired t - test same subject

test statistic = $\dfrac{\bar{d}-k}{\frac{S_d}{\sqrt{n}}}$

$\bar{d}$：样本差的均值
k：H_0
S_d：样本差的标准差

(P264)

(P214) confidence interval

反证法

e.g. 设 $\sqrt{2} = \frac{a}{b}$（a、b 互质）→ 设 $\sqrt{2}$ 为有理数

$\sqrt{2}b = a$

$2b^2 = a^2$

$2b^2 = (2c)^2$ → a 是偶数

$\cancel{2}b^2 = \cancel{4}^{2}c^2$

$b^2 = 2c^2$ → b 也是偶数

∴ a、b 不可能互质

∴ $\sqrt{2}$ 不是有理数

$\begin{cases} \text{type I error：弃真：放弃正确 } H_0 \\ \text{type II error：纳伪：接纳错误 } H_0 \end{cases}$

type II error 错误的认为 H_0 是对的，没有能够拒绝 H_0，H_0 实际上是不对的

5 5 3 = critical value

area = 1 − significance level μ：假设分布平均值 σ：$\frac{\sigma}{\sqrt{n}}$

5 5 2

upper bound：积分积到 critical value μ：true mean（唯一真分布的平均值） σ：$\frac{\sigma}{\sqrt{n}}$

type I error = significance level

不标准（原分布）

$\dfrac{\text{critical value} - \mu_a}{\frac{\sigma}{\sqrt{n}}}$ = 标准化后 critical value（area^{-1}）

[5.5.3] inverse N.

	原假设 H_0 为真	原假设 H_0 为假
拒绝 H_0	type I error (α)	正确决策 (1−β)
不拒绝 H_0	正确决策 (1−α)	type II error

e.g. ① $\dfrac{166.854-\mu_a}{\frac{25}{\sqrt{36}}} = -0.524401 \quad (\Phi^{-1}\,0.3)$

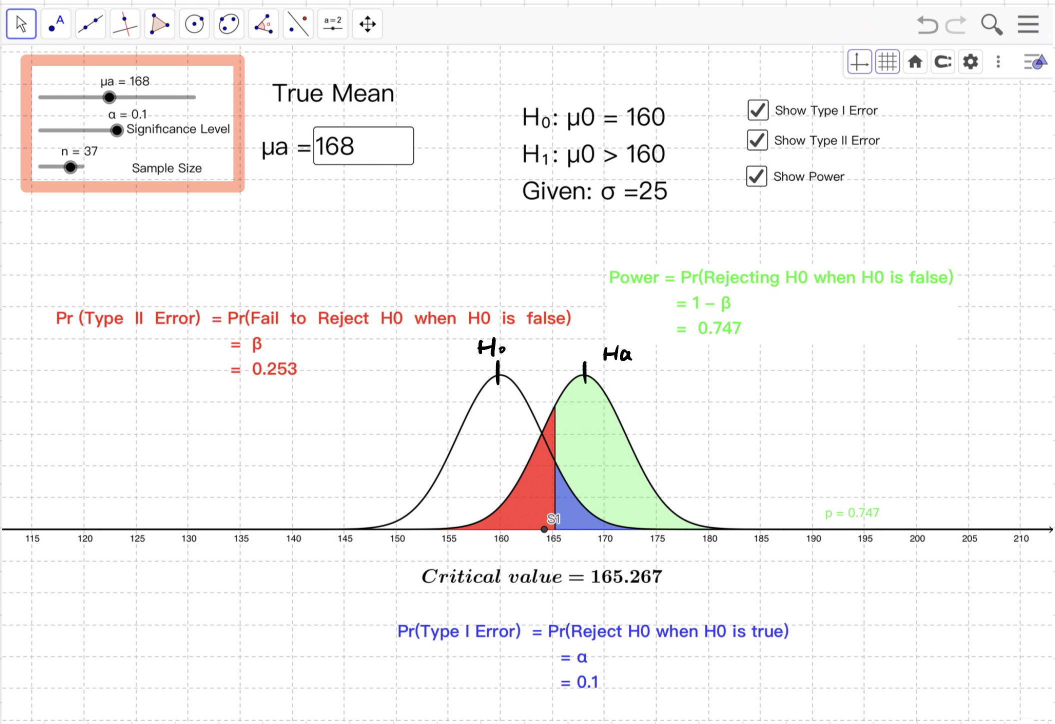
$\mu_a = 169.039$

② $\dfrac{166.854-\mu_a}{4.16667} = -0.253347$

$\mu_a = 167.91$

③ $\dfrac{166.854-\mu_a}{\frac{25}{\sqrt{36}}} = -0.582842$

$\mu_a = 169.283$



μ_a, α, n 对 β (type II error 的影响)

两个波峰距离越小，$\beta \uparrow$ $\quad |H_0 - H_a| \uparrow$, power $\uparrow$

$\alpha \downarrow$, $\beta \uparrow$

$n \downarrow$, $\beta \uparrow$

最小二乘法

$\Sigma(y_i - y_{pi})^2$ 越小越好　　y_i：实际值　y_{pi}：预测线上的值

平方和最小线 — 最小二乘回归线 least-square regression line (LSRL)

方差

$$\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n}$$

$$=\frac{\Sigma x_i^2+n\bar{x}^2-2\bar{x}\cdot n\bar{x}}{n}$$

$$=\frac{\Sigma x_i^2}{n}-(\bar{x})^2 \checkmark \qquad \left(\frac{\Sigma x}{n}\right)^2=\bar{x}^2$$

$$=E(x^2)-(E(x))^2$$

$$\frac{1^2+2^2+3^2+4^2+5^2+6^2}{6}-(3.5)^2=2.916667$$

$\sigma=\sqrt{2.916667}=1.71$　[表格 4, 1, 1]

$$\sum_{i=1}^{n}(a+bx_i-y_i)^2$$

$$=\sum_{i=1}^{n}(a^2+b^2x_i^2+y_i^2+2abx_i-2ay_i-2bx_iy_i)$$

$$=\sum_{i=1}^{n}a^2+b^2\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}y_i^2+2ab\sum_{i=1}^{n}x_i-2a\sum_{i=1}^{n}y_i-2b\sum_{i=1}^{n}x_iy_i$$

$$=n\times a^2+b^2\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}y_i^2+2abn\bar{x}-2an\bar{y}-2b\sum_{i=1}^{n}x_iy_i$$

$$=n\times a^2+2a\left(bn\bar{x}-n\bar{y}\right)+b^2\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}y_i^2-2b\sum_{i=1}^{n}x_iy_i$$

$$=n\left(a^2+2a\left(b\bar{x}-\bar{y}\right)+\left(b\bar{x}-\bar{y}\right)^2\right)-n\left(b\bar{x}-\bar{y}\right)^2+b^2\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}y_i^2-2b\sum_{i=1}^{n}x_iy_i$$

$$=n\left(a+b\bar{x}-\bar{y}\right)^2-n\left(b^2\left(\bar{x}\right)^2+\left(\bar{y}\right)^2-2b\bar{x}\bar{y}\right)+b^2\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}y_i^2-2b\sum_{i=1}^{n}x_iy_i$$

$$=n\left(a+b\bar{x}-\bar{y}\right)^2-nb^2\left(\bar{x}\right)^2-n\left(\bar{y}\right)^2+2nb\bar{x}\bar{y}+b^2\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}y_i^2-2b\sum_{i=1}^{n}x_iy_i$$

$$=\left(a+b\bar{x}-\bar{y}\right)^2+\left(\sum_{i=1}^{n}x_i^2-n\left(\bar{x}\right)^2\right)\times b^2+\left(2n\bar{x}\bar{y}-2\sum_{i=1}^{n}x_iy_i\right)\times b-n\left(\bar{y}\right)^2+\sum_{i=1}^{n}y_i^2$$

$$\begin{cases} b=-\dfrac{\left(2n\bar{x}\bar{y}-2\sum_{i=1}^{n}x_iy_i\right)}{2\left(\sum_{i=1}^{n}x_i^2-n\left(\bar{x}\right)^2\right)}=-\dfrac{n\bar{x}\bar{y}-\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2-n\left(\bar{x}\right)^2}=\dfrac{\sum_{i=1}^{n}\frac{x_iy_i}{n}-\bar{x}\bar{y}}{\sum_{i=1}^{n}\frac{x_i^2}{n}-\left(\bar{x}\right)^2}=\dfrac{\sum_{i=1}^{n}\frac{x_iy_i}{n}-\bar{x}\bar{y}}{\sum_{i=1}^{n}\frac{x_ix_i}{n}-\bar{x}\bar{x}} \\ a=\bar{y}-b\bar{x} \end{cases}$$

★ $y = \underbrace{\dfrac{\sum_{i=1}^{n} \frac{x_i y_i}{n} - \bar{x}\bar{y}}{\sum_{i=1}^{n} \frac{x_i x_i}{n} - \bar{x}\bar{x}}}_{b} \cdot x + (\bar{y} - b\bar{x})$ [表格 4, 1, 3]

$\bar{y} = b\bar{x} + (\bar{y} - b\bar{x})$

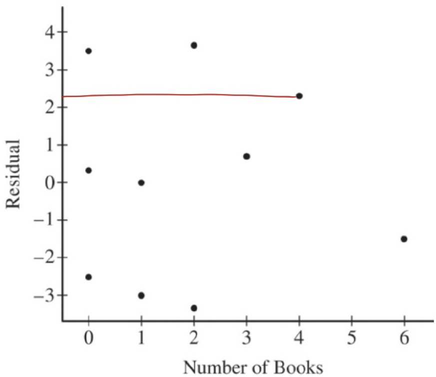
回归线一定经过 $(\bar{x}, \bar{y})$

(1, 1) (2, 3) (4, 10)

$$\underbrace{\dfrac{\frac{1\times1+2\times3+4\times10}{3} - \left(\frac{1+2+4}{3}\right)\cdot\left(\frac{1+3+10}{3}\right)}{\frac{1\times1+2\times2+4\times4}{3} - \left(\frac{1+2+4}{3}\right)\cdot\left(\frac{1+2+4}{3}\right)}}_{b} x + \left(\underbrace{\frac{1+3+10}{3}}_{y} - 3.07 \cdot \frac{1+2+4}{3}\right)$$

$= 3.07x - 2.5$

32. The weight, in pounds, of a full backpack and the corresponding number of books in the backpack were recorded for each of 10 college students. The resulting data were used to create the residual plot and the regression output shown below.



$y = a + bx = 10.53 + 0.53x = 10.53 + 0.53 \cdot 4 = 12.65$ $12.65 + 2.2 = 14.65$

	Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
截距 a	Intercept	10.53	1.23	≠ 0	8	8.57	< 0.0001
slope b	Slope	0.53	0.46	≠ 0	8	1.15	0.2825

Which of the following values is closest to the actual weight, in pounds, of the backpack for the student who had 4 books in the backpack?

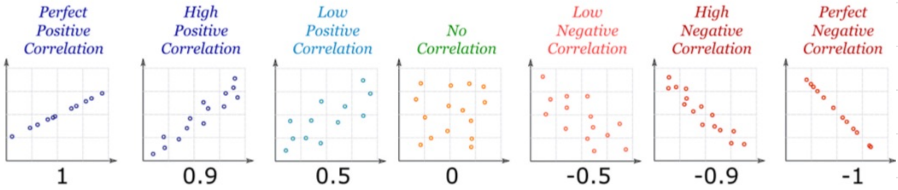
(A) 8
(B) 10
(C) 13
(D) 15
(E) 17

r 相关系数

$-1 < r < 1$

点和线的贴近程度

A correlation is assumed to be **linear** (following a line).



$$r = \frac{1}{n-1} \Sigma \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{x}}{S_y} \right) = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2 \Sigma (y_i - \bar{y})^2}}$$ 无单位

$y = a + \underline{b}x$ = 乘回归线的斜率 / 陡峭程度

$b \downarrow$ slope

SST (total variability) − SSE (unexplained) = SSR (explained)

SSE: 残余误差

SSE + SSR = SST

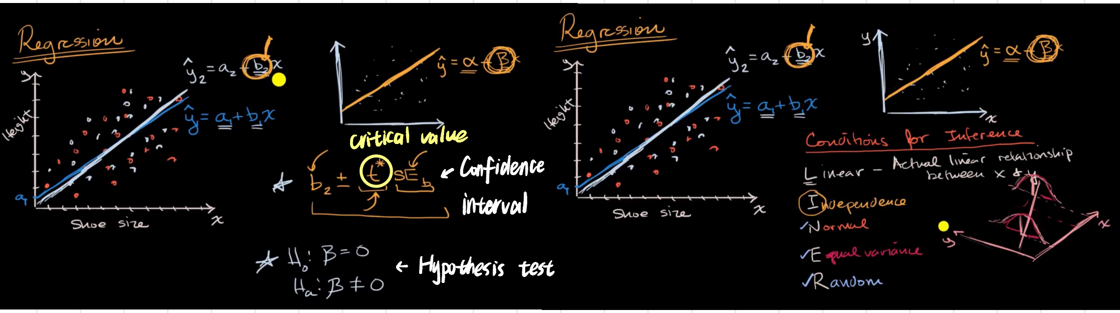
$r^2 = \frac{SSR}{SST}$ proportion of the variation that can be explained

SSE = 0 ⇒ SST = SSR ⇒ 相关性最强

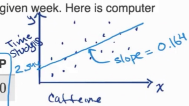
$$\frac{\sqrt{\Sigma (y_i - \bar{y})^2}}{\sqrt{\Sigma (x_i - \bar{x})^2}} = \frac{S_y}{S_x}$$ 计算式

$b = r \cdot \frac{S_y}{S_x}$ b, r 关系式

最小二乘回归线的置信区间和假设检验



Musa is interested in the relationship between hours spent studying and caffeine consumption among students at his school. He randomly selects 20 students at his school and records their caffeine intake (mg) and the amount of time spent studying in a given week. Here is computer output from a least-squares regression analysis on his sample:



Predictor	Coef	SE Coef	T	P
Constant	$\alpha =$ 2.544	0.134	18.955	0.000
Caffeine	$\beta =$ 0.164	0.057	2.862	0.010

$S = 1.532$ R-sq $= 60.0\%$ SE t^* P-value

Assume that all conditions for inference have been met.

What is the 95% confidence interval for the slope of the least squares regression line?

$0.164 \pm t^* \cdot 0.057$

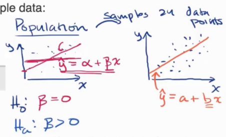
$df = 20 - 2 = 18$

khanacademy.org

Jian obtained a random sample of data on how long it took each of 24 students to complete a timed reaction game and a timed memory game. He noticed a positive linear relationship between the times on each task. Here is computer output on the sample data:

Summary statistics

Variable	n	Mean	StDev	SE Mean
x = reaction time	24	0.398	0.133	0.027
y = memory time	24	43.042	8.554	1.746



$H_0: \beta = 0$

$H_a: \beta > 0$

Regression: memory vs. reaction

Predictor	Coef	SE Coef
Constant	$a =$ 37.200	5.579
Reaction	$b =$ 14.686	13.329

$S = 8.515$ R-sq $= 5.2\%$

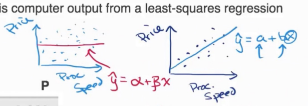
Assume that all conditions for inference have been met.

Calculate the test statistic that should be used for testing a null hypothesis that the population slope is actually 0?

$z = \dfrac{b - \beta_0}{\sigma_b}$ $t = \dfrac{b - \beta_0}{SE} = \dfrac{14.686 - 0}{13.329}$

khanacademy.org

Alicia took a random sample of mobile phones and found a positive linear relationship between their processor speeds and their prices. Here is computer output from a least-squares regression analysis on her sample:



Regression: Price vs. speed

Predictor	Coef	SE Coef	T	P
Constant	127.092	57.507	2.210	0.032
Speed	6.084	2.029	2.999	0.004

P-value $= P(t \geq 2.999) = 0.002$

Alicia wants to test $H_0: \beta = 0$ vs. $H_a: \beta > 0$. Assume that all conditions for inference have been met.

$0.002 < 0.010 \Rightarrow$ reject H_0

At the $\alpha = 0.01$ level of significance, is there sufficient evidence to conclude a positive linear relationship between these variables for all mobile phones? Why?

Yes, because P-value $< \alpha$ ⇒ reject H_0 ⇒ suggests

khanacademy.org

Hashem obtained a random sample of students and noticed a positive linear relationship between their ages and their backpack weights. A 95% confidence interval for the slope of the regression line was 0.39 ± 0.23.

Hashem wants to use this interval to test $H_0: \beta = 0$ vs. $H_a: \beta \neq 0$ at the $\alpha = 0.05$ level of significance. Assume that all conditions for inference have been met.

95% confidence interval: $[0.16, 0.62]$

Assuming H_0 true, we are in $\leq 5\%$ situations where β not overlap with 95% interval

⇓

reject H_0 ⇒ suggest $H_a: \beta \neq 0$

⇓

there is a non-zero linear relationship between ages & backpack weights.

khanacademy.org

卡方分布 (inference for categorical data)

独立性检验 & 同质性检验

H_0: independent
H_a: not independent

association { independence / homogeneity → proportion

$df = (r-1)(c-1)$ $TS = \chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i}$ expected value $= \frac{R_i \cdot C_i}{n}$

① 菜单 7.1,1 control var. (↓ store) enter 输入矩阵

② 菜单 6,7,8

expected value (4): var. expect.

拟合优度检验

有一标准分布,检验抽样是否符合

$df = k - 1$

expected value

电子表格 再加一列,=, var. pro, × 总数, enter

电子表格 4.4.7 compare x^2 with $\underbrace{\text{critical value}}$ 5.5.9 卡方检验都是单尾

in all conditions, $E_i \geq 5$, then $\Sigma \left(\frac{(O_i - E_i)^2}{E_i}\right) \sim x^2$ is a good approximation.

sample size 一般小于等于总体的 10% → $n \leq \frac{1}{10} N$

H_0: 两者一致 (TS 小, P value 大)
H_a: 两者不一致 (TS 大, P value 小)

Ha对应小概率事件 ⇄ 大χ^2值 ⇄ 变量比例高度不一致

$$\begin{bmatrix} 60 & 20 \\ 60 & 30 \\ 60 & 90 \end{bmatrix}$$

$\chi^2 = 31.5$

P value $= 1.45 \times 10^{-7}$

拒绝H_0，接受Ha　一定不独立

$$\begin{bmatrix} 60 & 20 \\ 60 & 20 \\ 60 & 20 \end{bmatrix}$$

$\chi^2 = 0$

P value $= 1$

无法拒绝H_0　可能独立

希望小概率事件发生，以此拒绝H_0，接受Ha.

在H_0的条件下，

- 小概率事件一旦发生，说明H_0错误 → 所以Ha正确 (P value < 0.05)
- 小概率事件不发生，H_0有可能是对的. → 无法拒绝H_0 (P value > 0.05)

在卡方检验中，H_0表示相类似 / 无差异 / 同比例 / 相互独立.
no association

所有 P value 都是基于H_0成立

P value：H_0正确的条件下，其他更极端事件发生概率

	数学好	数学不好
左	70	30
右	30	70

$\chi^2 = 32$

$$\left(\frac{(70-50)^2}{50} + \frac{(30-50)^2}{50}\right) \times 2$$

$$(8+8) \times 2 = 32$$

70	30
70	30

$\chi^2 = 0$

统计问题定性分析

1. confidence interval vs. hypothesis test (prediction / Chi-square)
   $a \le x \le b$ statement with $\le$, $\geq$, $=$

2. proportion vs. mean
   - 男生人数占比（女生人数占比） · 身高
   · 体重
   - 得新冠的患病率 · 血压

3. independent vs. dependent
   (unpaired vs. paired)

4. one sample vs. two samples

5. categorical vs. quantitative

In 95% of all sample condition, the method would yeild an interval captures the true parameter value.

1、笔记
2、flashcards quizelet
3、统计问题定性分析
http://www.ltcconline.net/greenl/java/Statistics/catStatProb/categorizingStatProblemsJavaScript.html
4、定量计算（TI-Nspire）
5、官方单元真题
6、往年真题
7、必考 ① 双变量
② 概率
★ ③ confidence interval
★ ④ 假设检验
⑤ 卡方 / slope